Predicting West Nile Virus Mosquito Positivity Rates and Abundance: A Comparative Evaluation of Machine Learning Methods for Epidemiological Applications

Julianna Schneider[1], Alessandro Greco[2], Jillian Chang[2], Maria Molchanova[2], and Luke Shao[2]

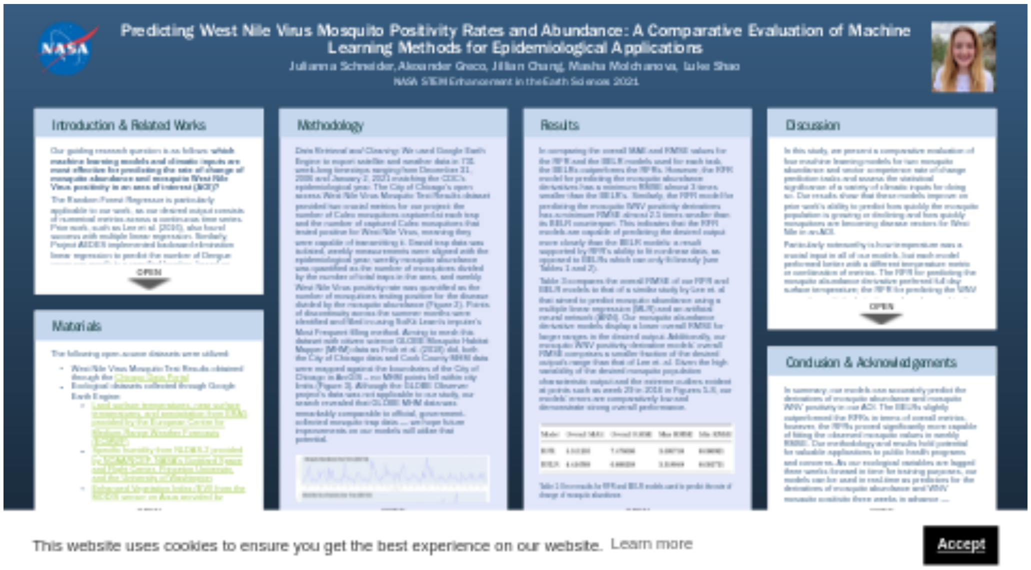
[1]NASA SEES (STEM Enhancement in the Earth Sciences)
[2]NASA SEES (STEM Enhancement in Earth Sciences)

November 23, 2022

Abstract

Mosquitoes are major vectors of disease and thus a key public health concern. Some cities have programs to track mosquito abundance and vector competence, but such fieldwork is expensive, time-consuming, and retrospective. We present a comparative analysis of two machine-learning-based regression techniques for forecasting the rate at which mosquito abundance changes and the rate at which mosquitoes test positive for West Nile Virus (WNV) in our AOI, the City of Chicago, three weeks in advance. We selected an initial pool of climatic inputs based on the findings of prior work. Ordinary least squares regression was run on each input individually and then in various groups. A p-value cutoff of 0.05 was used to determine which were best suited for predicting the derivatives of mosquito abundance and WNV positivity rate. Using these inputs, we trained four machine learning models using two types of regression: a Random Forest Regressor (RFR) and Backward Elimination Linear Regression (BELR). We optimized our RFR's hyperparameters using Randomized Search Cross Validation and further reduced our BELR inputs using a p-value of 0.05. The enhanced vegetation index and temperature, described in various metrics, emerged as common inputs across the four models. In three of the four models, the respective temperature metric was the most important feature while EVI varied between second and last place. Our root mean square error largely resided within the hundredths place or less, but spiked at novel, week-to-week extremes in the testing data. Our methodology and results indicate valuable directions for future research into forecasting mosquito population abundance and vector competence. This work is particularly applicable to public health programs, as our models' use of open-source, remote sensing data to predict how quickly the mosquito population and their vector competence will change three weeks in advance streamlines disease monitoring and prevention.

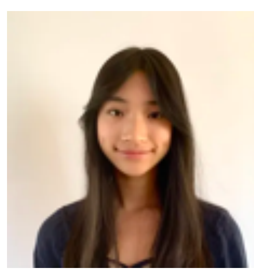
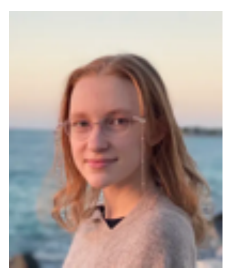
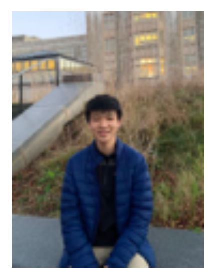
Predicting West Nile Virus Mosquito Positivity Rates and Abundance: A Comparative Evaluation of Machine Learning Methods for Epidemiological Applications



Julianna Schneider, Alexander Greco, Jillian Chang, Masha Molchanova, Luke Shao

NASA STEM Enhancement in the Earth Sciences 2021



PRESENTED AT:

INTRODUCTION & RELATED WORKS

Our guiding research question is as follows: **which machine learning models and climatic inputs are most effective for predicting the rate of change of mosquito abundance and mosquito West Nile Virus positivity in an area of interest (AOI)?**

The Random Forest Regressor is particularly applicable to our work, as our desired output consists of numerical metrics across a continuous time series. Prior work, such as Lee et al. (2016), also found success with multiple linear regression. Similarly, Project AEDES implemented backward elimination linear regression to predict the number of Dengue cases per month in a specified location, based on weather variables (temperature and rainfall) and google search trends (Ligot et. al, 2021). Hence, we evaluated the performance of a Random Forest Regressor (RFR) and Backward Elimination Linear Regression (BELR) for each prediction task.

Our selection of precipitation, temperature, humidity, and vegetation metrics as model inputs was informed by the success of prior work. Francisco et al. (2021) used monthly average precipitation, average land surface temperature, and flood susceptibility data to prove a significant correlation between precipitation and dengue outbreaks at a one-month lag in Manila, Philippines. Hassan et al. (2012) derived environmental variables such as urbanization level, Land Use Land Cover, Normalized Difference Vegetation Index (NDVI) from Landsat TM5 and Ikonosimageries to characterize landscape features likely associated with mosquito breeding habitats in Cairo, Egypt; land cover type and vegetation proved important indicators of potential mosquito habitats. Früh et al. (2018) trained a variety of machine learning models on citizen science data to predict the occurrence of Aedes japonicus japonicus, an invasive mosquito species in Germany. Their work indicated that mean precipitation, mean temperature, and drought index were the most accurate predictors of mosquito occurrence. Chen et al. (2019) indicated that landscape factors alone yield equal or more accurate modeling when compared to or paired with socioeconomic factors. Consequently, we evaluated the performance of a variety of machine learning regressors powered by open-source data describing the aforementioned ecological factors.

MATERIALS

The following open-source datasets were utilized:

- West Nile Virus Mosquito Test Results obtained through the Chicago Data Portal (https://data.cityofchicago.org/Health-Human-Services/West-Nile-Virus-WNV-Mosquito-Test-Results/jqe8-8r6s)
- Ecological datasets collected through Google Earth Engine:
  - Land surface temperatures, near surface temperatures, and precipitation from ERA5 provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) (https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_HOURLY?hl=en)
  - Specific humidity from NLDAS-2 provided by NOAA/NCEP, NASA's Goddard Space and Flight Center, Princeton University, and the University of Washington (https://developers.google.com/earth-engine/datasets/catalog/NASA_NLDAS_FORA0125_H002?hl=en)
  - Enhanced Vegetation Index (EVI) from the MODIS sensor on Aqua provided by Google and NASA's EOSDIS (https://developers.google.com/earth-engine/datasets/catalog/MODIS_MYD09GA_006_EVI)

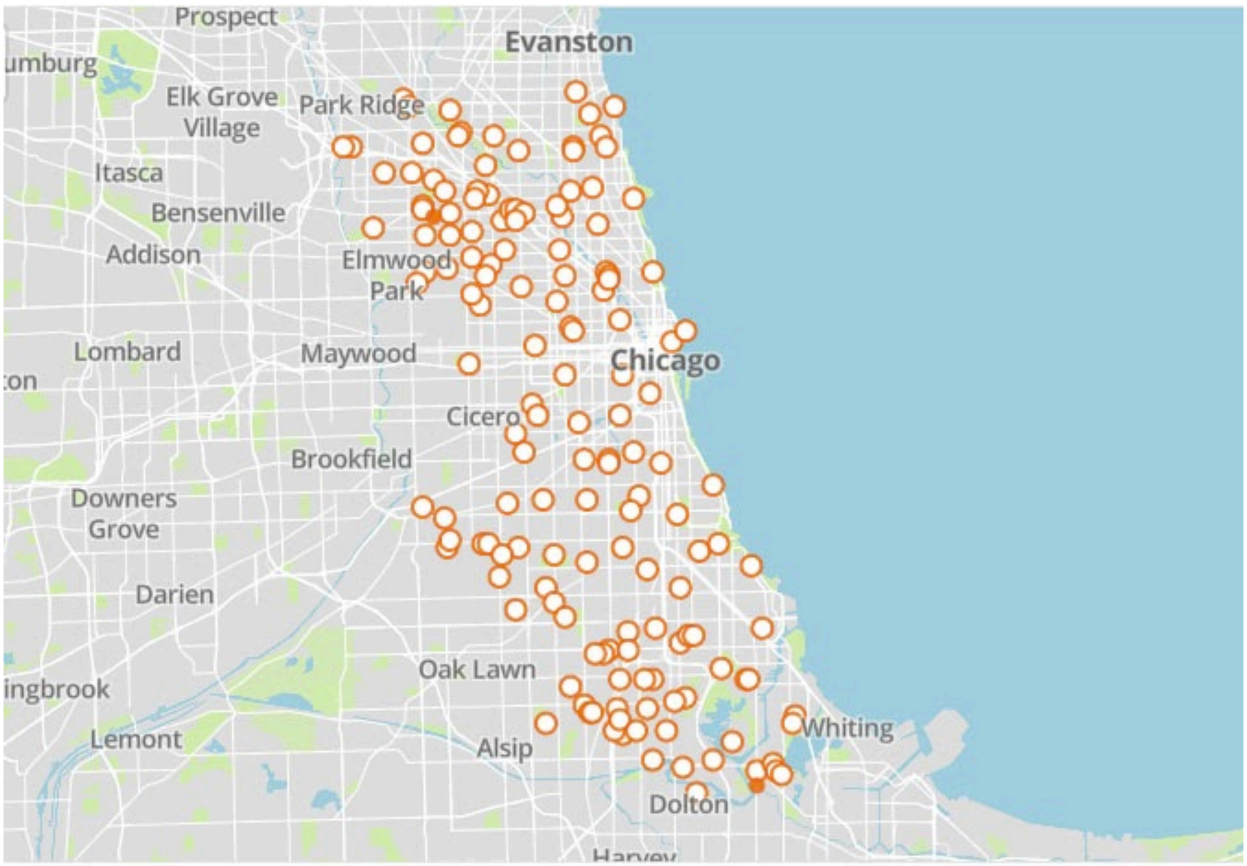


Figure 1: A visualization of the measurement locations in the Chicago Data Portal's dataset.

METHODOLOGY

Data Retrieval and Cleaning: We used Google Earth Engine to export satellite and weather data in 731 week-long timesteps ranging from December 31, 2006 and January 2, 2021 matching the CDC's epidemiological year. The City of Chicago's open access West Nile Virus Mosquito Test Results dataset provided two crucial metrics for our project: the number of Culex mosquitoes captured at each trap and the number of captured Culex mosquitoes that tested positive for West Nile Virus, meaning they were capable of transmitting it. Gravid trap data was isolated, weekly measurements were aligned with the epidemiological year, weekly mosquito abundance was quantified as the number of mosquitoes divided by the number of total traps in the area, and weekly West Nile Virus positivity rate was quantified as the number of mosquitoes testing positive for the disease divided by the mosquito abundance (Figure 2). Points of discontinuity across the summer months were identified and filled in using SciKit Learn's imputer's Most Frequent filling method. Aiming to mesh this dataset with citizen science GLOBE Mosquito Habitat Mapper (MHM) data as Früh et al. (2018) did, both the City of Chicago data and Cook County MHM data were mapped against the boundaries of the City of Chicago in ArcGIS – no MHM points fell within city limits (Figure 3). Although the GLOBE Observer project's data was not applicable to our study, our search revealed that GLOBE MHM data was remarkably comparable to official, government-collected mosquito trap data — we hope future improvements on our models will utilize that potential.

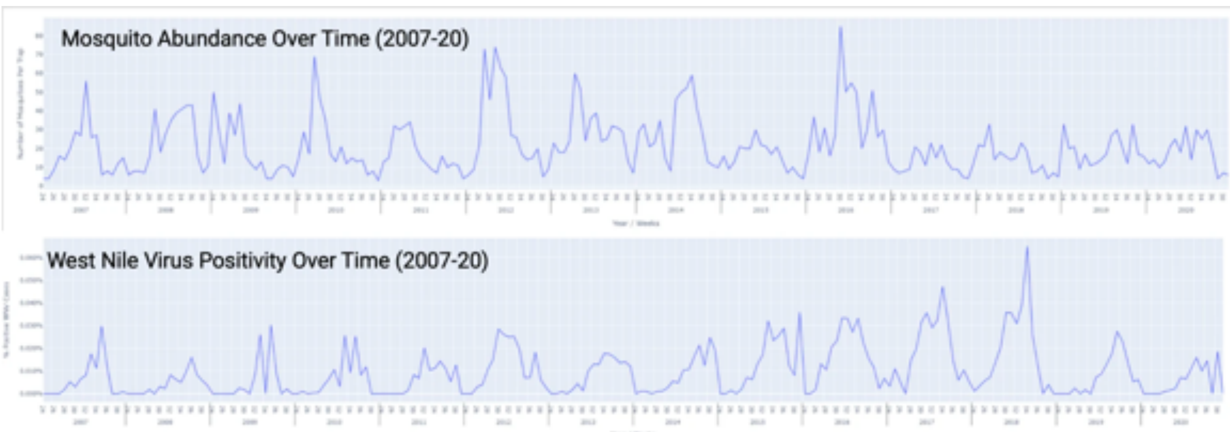


Figure 2: Mosquito abundance and mosquito WNV positivity graphed over time.

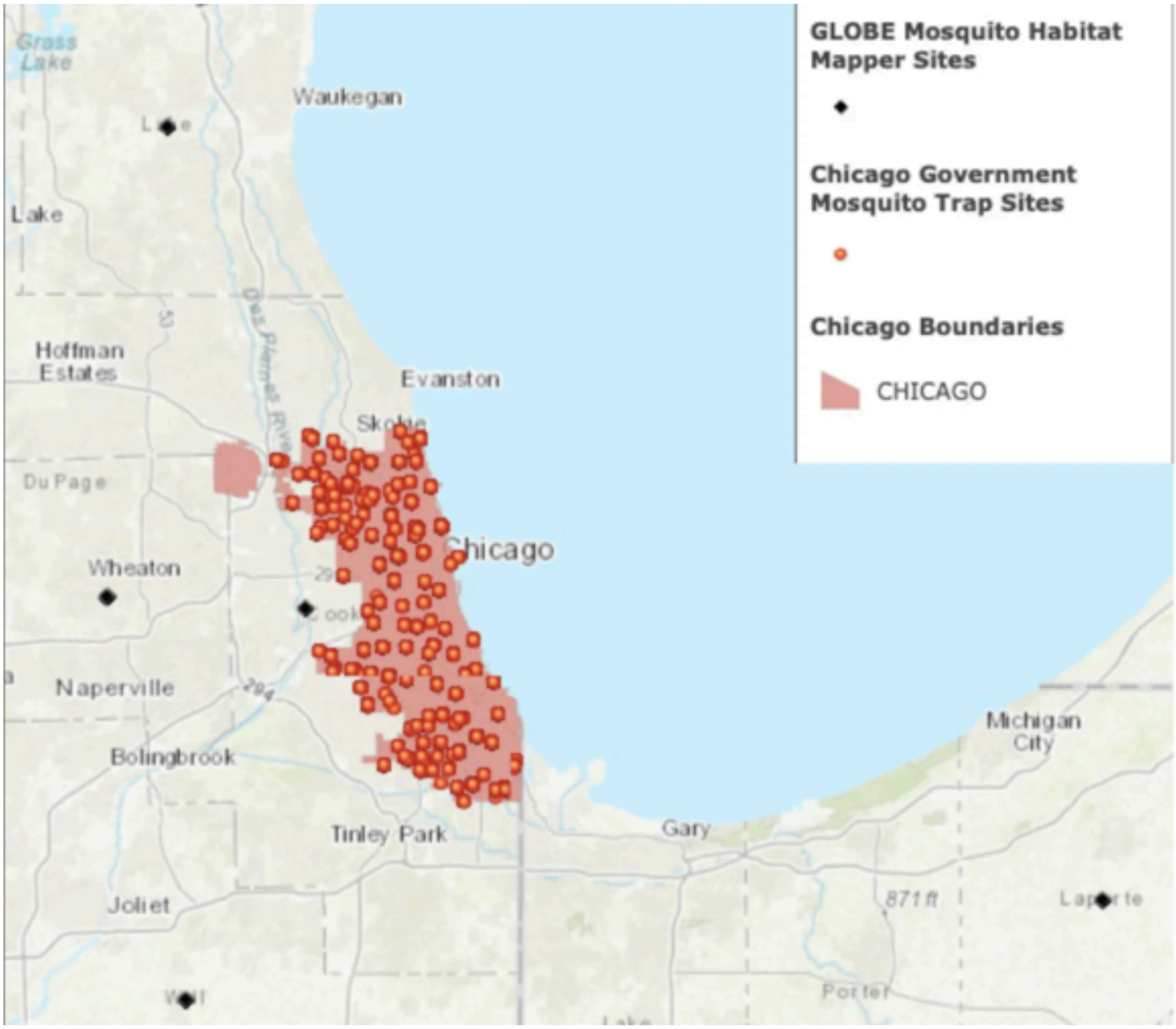


Figure 3: GLOBE MHM sites and Chicago government mosquito trap sites plotted in ARCGis.

Pre-processing: Lopez et al. (2014) observed higher correlations between dengue outbreaks and environmental factors when time lags were introduced. Inspired by this work, we examined the relationship between the various climatic variables collected from our literature review and the mosquito abundance and positivity outputs by graphing them. Shifting EVI, Land Surface Temp, and Specific Humidity (as it relates to mosquito abundance), and total precipitation (as it relates to West Nile Virus positivity rates) three weeks forward in time better aligned the input and output peaks, as seen in Figure 4. We then padded our dataset, extending the weeks in each summer to 21-41, so as to calculate the derivative of mosquito abundance and WNV positivity for our weeks of interest, 22-40.

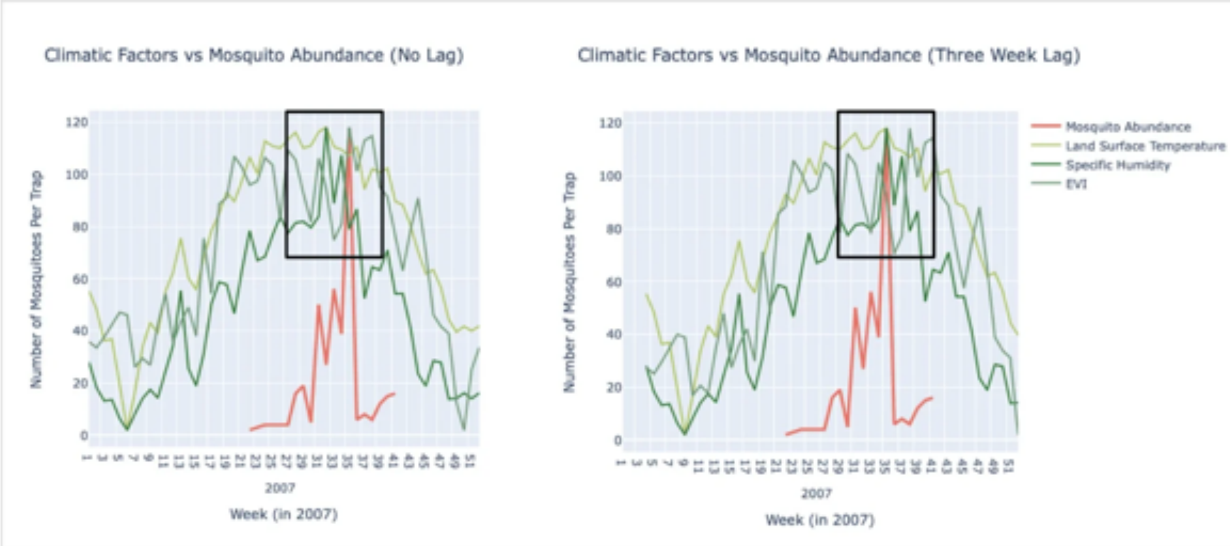


Figure 4: Introducing a three-week lag to our climatic variables to help their peaks align more closely with the peaks in mosquito abundance - same technique used for positivity.

Feature Selection and Model Training: Having assembled our initial pool of independent climatic variables based on the findings of prior work, we narrowed down our pool of inputs using ordinary least squares (OLS) regression. First, we ran OLS regression on each input individually to establish which had statistically significant correlations with mosquito abundance and which had statistically significant correlations with mosquito WNV positivity using a p-value of 0.05. Then, we grouped promising climatic inputs for each prediction task into various sets and ran OLS regression on each set, revealing EVI, land surface temperature, total precipitation, and specific humidity as the optimal inputs for predicting mosquito abundance and EVI, specific humidity, near-surface temperature range, and night-time temperature as the optimal inputs for predicting mosquito WNV positivity. We then trained an RFR and BELR to predict the rate of change of mosquito abundance and an RFR and BELR to predict the rate of change of mosquito WNV positivity. The RFRs were built using SciKit-Learn's RFR model and trained using its Randomized Search Cross Validation tool. On running 100 iterations with three cross folds each, the optimal hyperparameters for each RFR emerged. The BELRs were built using Sci-Kit Learn's Linear Regression model and inputs were eliminated using OLS regression to determine which inputs were statistically insignificant to the BELR's predictions.

RESULTS

In comparing the overall MAE and RMSE values for the RFR and the BELR models used for each task, the BELRs outperforms the RFRs. However, the RFR model for predicting the mosquito abundance derivatives has a minimum RMSE almost 3 times smaller than the BELR's. Similarly, the RFR model for predicting the mosquito WNV positivity derivatives has a minimum RMSE almost 2.5 times smaller than its BELR counterpart. This indicates that the RFR models are capable of predicting the desired output more closely than the BELR models: a result supported by RFR's ability to fit nonlinear data, as opposed to BELRs which can only fit linearly (see Tables 1 and 2).

Table 3 compares the overall RMSE of our RFR and BELR models to that of a similar study by Lee et. al that aimed to predict mosquito abundance using a multiple linear regression (MLR) and an artificial neural network (ANN). Our mosquito abundance derivative models display a lower overall RMSE for larger ranges in the desired output. Additionally, our mosquito WNV positivity derivative models' overall RMSE comprises a smaller fraction of the desired output's range than that of Lee et. al. Given the high variability of the desired mosquito population characteristic output and the extreme outliers evident at points such as week 29 in 2016 in Figures 5-8, our models' errors are comparatively low and demonstrate strong overall performance.

Model	Overall MAE	Overall RMSE	Max RMSE	Min RMSE
RFR	5.311205	7.476696	3.699710	0.000921
BELR	4.454769	6.696208	3.318949	0.002721

Table 1: Error results for RFR and BELR models used to predict the rate of change of mosquito abundance.

Model	Overall MAE	Overall RMSE	Max RMSE	Min RMSE
RFR	0.004442	0.006522	0.002433	1.378654e-06
BELR	0.004279	0.006451	0.002520	3.539568e-06

Table 2: Error results for RFR and BELR models used to predict the rate of change of mosquito WNV positivity rates.

Model	Overall RMSE	Range of Desired Output
RFR for Mosquito Abundance Derivative	7.476696	98
BELR for Mosquito Abundance Derivative	6.696209	98
RFR for Mosquito WNV Positivity	0.006522	0.0435
BELR for Mosquito WNV Positivity	0.006451	0.0435
MLR for Mosquito Abundance	17.53	78
ANN for Mosquito Abundance	14.38	78

Table 3: Comparison of overall RMSEs for our RFR and BELR models and Lee et. al's MLR and ANN.

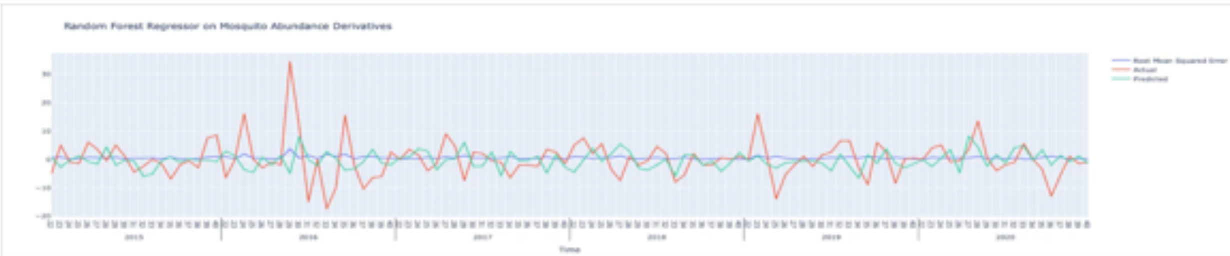


Figure 5: RFR results for predicting the rate of change of mosquito abundance.

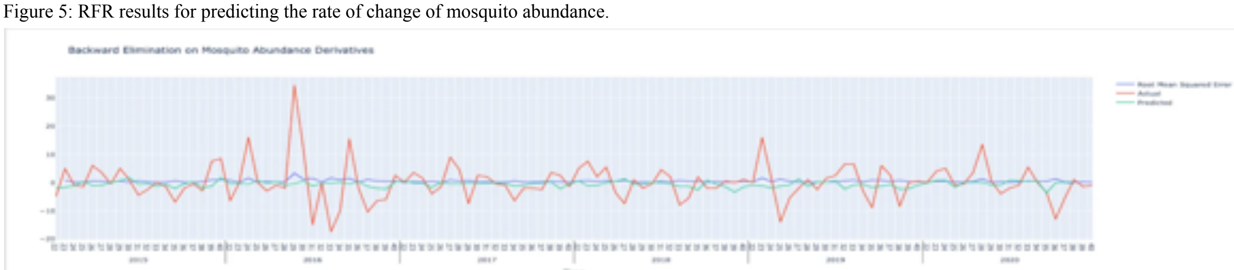


Figure 6: BELR results for predicting the rate of change of mosquito abundance.

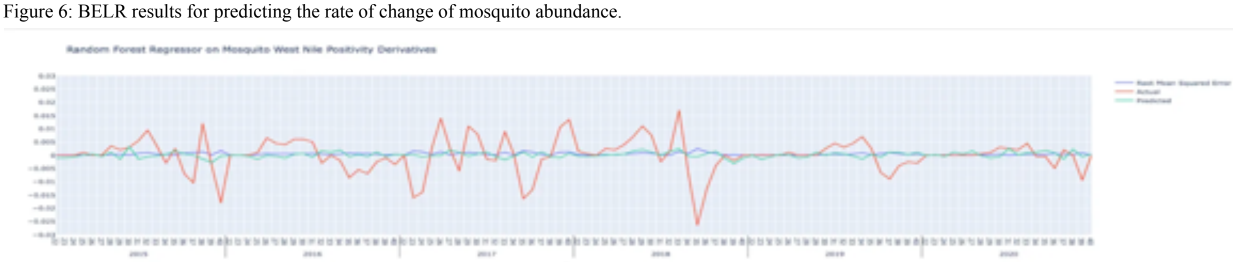


Figure 7: RFR results for predicting the rate of change of mosquito WNV positivity.

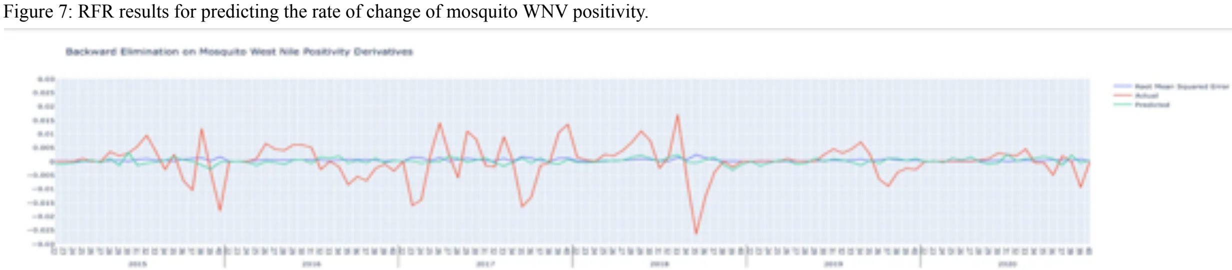


Figure 8: BELR results for predicting the rate of change of mosquito WNV positivity.

DISCUSSION

In this study, we present a comparative evaluation of four machine learning models for two mosquito abundance and vector competence rate of change prediction tasks and assess the statistical significance of a variety of climatic inputs for doing so. Our results show that these models improve on prior work's ability to predict how quickly the mosquito population is growing or declining and how quickly mosquitoes are becoming disease vectors for West Nile in an AOI.

Particularly noteworthy is how temperature was a crucial input in all of our models, but each model performed better with a different temperature metric or combination of metrics. The RFR for predicting the mosquito abundance derivative preferred full day surface temperature; the RFR for predicting the WNV mosquito positivity derivative preferred a combination of full day near surface temperature and full day land surface night temperatures; the backward elimination model for the abundance derivative preferred full day surface temperature; and the backward elimination model for positivity preferred land surface night temperatures alone.

Unlike much of the literature that informed our research, precipitation did not prove a significant factor across our machine learning models. However, indirect measurements of water quantity, such as EVI, did prove crucial and common across all models. This may be the result of differences between our AOI and that of other studies or the OLS regression we used to narrow down our climatic inputs, which only fits — and therefore deems significant — linear correlations.

Our findings provide avenues for further research and a deeper understanding of how mosquito populations thrive and become more potent disease vectors in response to climatic variation (Figure 9). Similarly, there remain areas for improvement upon our research. First, we applied our methodology to a single area of interest — to test its robustness, future work should examine how well the development procedure adjusts to different areas of interest. Second, we averaged data over the entirety of Chicago, making our predictions applicable to the whole of Chicago but not specific to a single area within it. With more consistent and detailed data recorded on more frequent time steps, our model would likely perform better and output predictions further localized to mosquito and West Nile hotspots within the greater City of Chicago.

CONCLUSION & ACKNOWLEDGEMENTS

In summary, our models can accurately predict the derivatives of mosquito abundance and mosquito WNV positivity in our AOI. The BELRs slightly outperformed the RFRs in terms of overall metrics, however, the RFRs proved significantly more capable of fitting the observed mosquito values in weekly RMSE. Our methodology and results hold potential for valuable applications to public health programs and concerns. As our ecological variables are lagged three weeks forward in time for training purposes, our models can be used in real-time as predictors for the derivatives of mosquito abundance and WNV mosquito positivity three weeks in advance — providing public health officials with critical information on the development of mosquito populations in time for appropriate intervention and mitigation as seen in Figure 9. Avenues for future work and development on our results revolve around how to further increase accuracy and best incorporate our methodology into existing public health initiatives.

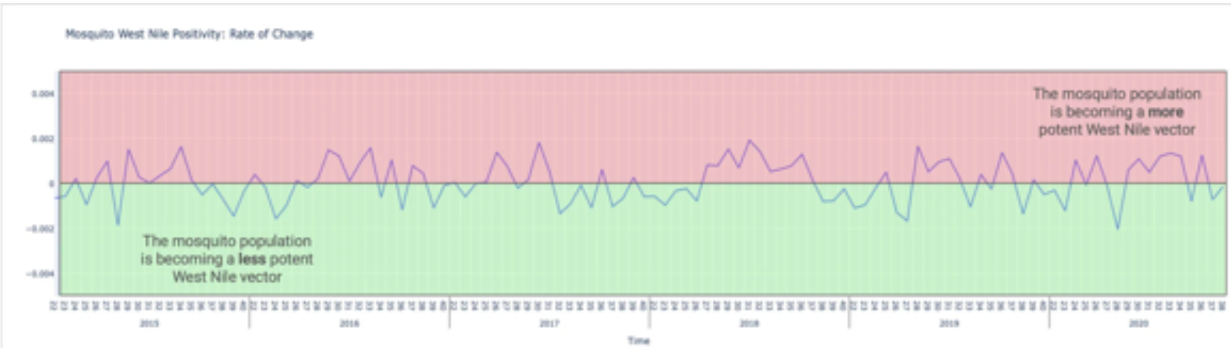


Figure 9: Annotated graph of Random Forest Regressor's predictions of the rate of change of mosquiot WNV positivity rate illustrating the public health program applicability of our findings.

Acknowledgements:

Our gratitude goes to our NASA SEES 2021 mentors Dr. Erika Podest, Dr. Rusanne Low, Ms. Cassie Soeffing, Mr. Peder Nelson, and Dr. Becky Boger!

The material contained in this poster is based upon work supported by National Aeronautics and Space Administration (NASA) cooperative agreements NNX16AE28A to the Institute for Global Environmental Strategies (IGES) for the NASA Earth Science Education Collaborative (NESEC) and NNX16AB89A to the University of Texas Austin for the STEM Enhancement in Earth Science (SEES). Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NASA.

ABSTRACT

Mosquitoes are major vectors of disease and thus a key public health concern. Some cities have programs to track mosquito abundance and vector competence, but such fieldwork is expensive, time-consuming, and retrospective. We present a comparative analysis of two machine-learning-based regression techniques for forecasting the rate at which mosquito abundance changes and the rate at which mosquitoes test positive for West Nile Virus (WNV) in our AOI, the City of Chicago, three weeks in advance. We selected an initial pool of climatic inputs based on the findings of prior work. Ordinary least squares regression was run on each input individually and then in various groups to determine which were best suited for predicting the derivatives of mosquito abundance and WNV positivity rate. Using these inputs, we trained four machine learning models using two types of regression: a Random Forest Regressor (RFR) and Backward Elimination Linear Regression (BELR). We optimized our RFR's hyperparameters using Randomized Search Cross Validation and further reduced our BELR inputs using a p-value of 0.05. The enhanced vegetation index and temperature, described in various metrics, emerged as common inputs across the four models. In three of the four models, the respective temperature metric was the most important feature, while EVI varied between second and last place. Our root mean square error largely resided within the hundredths place or less, but spiked at novel, week-to-week extremes in the testing data. Our methodology and results indicate valuable directions for future research into forecasting mosquito population abundance and vector competence. This work is particularly applicable to public health programs, as our models' use of open-source, remote sensing data to predict how quickly the mosquito population and their vector competence will change three weeks in advance streamlines disease monitoring and prevention.

REFERENCES

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing, 114, 24–31.

Chai, T., & Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. Geoscientific model development, 7(3), 1247–1250.

Chen, S., Whiteman, A., Li, A., Rapp, T., Delmelle, E., Chen, G., Brown, C. L., Robinson, P., Coffman, M. J., Janies, D., et al. (2019). An operational machine learning approach to predict mosquito abundance based on socioeconomic and landscape patterns. Landscape Ecology, 34(6), 1295–1311.

Ding, F., Fu, J., Jiang, D., Hao, M., & Lin, G. (2018). Mapping the spatial distribution of aedes aegypti and aedes albopictus. Acta tropica, 178, 155–162.

Francisco, M. E., Carvajal, T. M., Ryo, M., Nukazawa, K., Amalin, D. M., & Watanabe, K. (2021). Dengue disease dynamics are modulated by the combined influences of precipitation and landscape: A machine learning approach. Science of The Total Environment, 148406.

Früh, L., Kampen, H., Kerkow, A., Schaub, G. A., Walther, D., & Wieland, R. (2018). Modelling the potential distribution of an invasive mosquito species: Comparative evaluation of four machine learning methods and their combinations. Ecological Modelling, 388, 136–144.

GLOBE. (2021). Global learning and observations to benefit the environment (globe) program. globe.gov

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. Remote sensing of Environment, 202, 18–27.

Hassan, A. N., El Nogoumy, N., & Kassem, H. A. (2013). Characterization of landscape features associated with mosquito breeding in urban cairo using remote sensing. The Egyptian Journal of Remote Sensing and Space Science, 16 (1), 63–69.

Koolhof, I. S., Gibney, K. B., Bettiol, S., Charleston, M., Wiethoelter, A., Arnold, A.-L., Campbell, P. T., Neville, P. J., Aung, P., Shiga, T., et al. (2020). The forecasting of dynamical ross river virus outbreaks: Victoria, australia. Epidemics, 30, 100377.

Lee, K. Y., Chung, N., & Hwang, S. (2016). Application of an artificial neural network (ann) model for predicting mosquito abundances in urban areas. Ecological Informatics, 36, 172–180.

Ligot, D., Toledo, M., & Melendres, R. (2021). Project aedes dpg repository wiki. https://github.com/Cirrolytix/aedes_dpg/wiki

Luman, D., Tweddale, T., Bahnsen, B., & Willis, P. (2004). Illinois land cover: Champaign, il, illinois state geological survey, illinois map 12, scale 1:500,000. https://files.isgs.illinois.edu/sites/default/files/maps/statewide/imap12.pdf

Orkins. (2021). Orkin's 2021 top mosquito cities list. https://www.orkin.com/press-room/orkins-2021-top-mosquito-cities

Roberts, M. (2021). West nile virus (wnv) mosquito test results. https://data.cityofchicago.org/Health-Human-Services/West-Nile-Virus-WNV- Mosquito-Test-Results/jqe8-8r6s

US Department of Commerce, N. (2021). Annual precipitation rankings for chicago, illinois. https://www.weather.gov/lot/Annual_Precip_Rankings_Chicago